Research Article | Vol. 12, No. 9 / 1 Sep 2021 / Biomedical Optics Express 5423
Biomedical Optics EXPRESS

Wavefront sensor-less adaptive optics using deep reinforcement learning

EDUARD DURECH,[1,3,4] WILLIAM NEWBERRY,[1] JONAS FRANKE,[2] AND MARINKO V. SARUNIC[2]

[1]School of Engineering Science, 8888 University Dr., Burnaby, BC V5A 1S6, Canada
[2]Institute of Biomedical Optics, University of Lübeck, 23562 Luebeck, Germany
[3]edurech@sfu.ca
[4]msarunic@sfu.ca

Abstract: Image degradation due to wavefront aberrations can be corrected with adaptive optics (AO). In a typical AO configuration, the aberrations are measured directly using a Shack-Hartmann wavefront sensor and corrected with a deformable mirror in order to attain diffraction limited performance for the main imaging system. Wavefront sensor-less adaptive optics (SAO) uses the image information directly to determine the aberrations and provide guidance for shaping the deformable mirror, often iteratively. In this report, we present a Deep Reinforcement Learning (DRL) approach for SAO correction using a custom-built fluorescence confocal scanning laser microscope. The experimental results demonstrate the improved performance of the DRL approach relative to a Zernike Mode Hill Climbing algorithm for SAO.

1. Introduction

Adaptive optics (AO), originally developed for earth-based astronomical imaging, is increasingly finding applications in microscopy. The impact of AO for vision science is described in a recent review article [1]. Specialized confocal bio-microscopes used for *in vivo* imaging of the retina are called confocal Scanning Laser Ophthalmoscopes (cSLO), which uses the refractive elements of the eye (cornea and intraocular lens) as the objective lens. The resolution attainable with cSLO ultimately depends on the optical quality of the eye.

In a typical AO system, a deformable mirror (DM) is used to introduce counter-aberrations and flatten the wavefront to a constant phase in order to restore diffraction limited image quality. This is generally done by measuring the wavefront, for example with a Shack-Hartmann wavefront sensor (SH-WFS) [2], and using control algorithms to shape a DM to correct for the aberrations.

Wavefront sensors unfortunately do not work in all scenarios; for example, a sample with multiple reflecting/scattering surfaces could confound accurate wavefront measurement. An application where SH-WFS is challenging is small animal (e.g. mouse) retinal imaging [3–5]. In the mouse, the retina is thick relative to the length of the eye, creating multiple reflections of different aberrations and confusing wavefront sensors [6]. Wavefront sensors also require additional calibration for non-common path aberrations.

One approach to overcome the challenges of wavefront sensors is to use Sensor-less Adaptive Optics (SAO), in which the signal or image may be used to directly control the DM. SAO methods are based on the principle that changing the wavefront with a DM has some relationship with the chosen image quality metric and can be maximized. A challenge with this approach to aberration correction is that the wavefront itself is only indirectly measured by its effect on the image and an image quality metric must be created. This can also be regarded as a benefit where the optimization is done on the image itself and seeks to optimize the quality, with the assumption that a representative and desired image quality metric can be accurately modeled. SAO is also able to correct for non-common path aberrations as the control system uses the final signal as its input.

#427970 https://doi.org/10.1364/BOE.427970
Journal © 2021 Received 19 Apr 2021; revised 20 Jun 2021; accepted 22 Jun 2021; published 6 Aug 2021

Biomedical Optics EXPRESS

For SAO, optimization methods such as a Zernike Mode Hill Climbing (ZMHC) [7] exploit the convex interaction of a single-mode aberration to the image quality metric around the point of correction; in essence, a DM is used to counter-aberrate one mode by stepping through several coefficient values for this mode and plotting it against the measured image quality metric. Around the true correction, the relation is ideally parabolic and each mode can be isolated and independently optimized [8]. This assumes that each mode has an independent effect on the image quality, but in practice several recursive optimizations must be done to properly correct the aberration because the modes interact through the merit function [9]. Examples of extrema-seeking SAO methods applied to retinal imaging include improvements to the coordinate search approach for ZMHC [10] and the use of gradient descent [11–13].

Multi-dimensional function-fitting can also be used to create a virtual representation of the Zernike space, such as using the Data-based Online Nonlinear Extremum-seeker (DONE) algorithm [14,15]. DONE anneals the influence of measurements over time so older measurements do not influence the fit as much as newer ones, which can be desirable due to target motion or blinking.

Some of the difficulties of SAO include the time to correct for aberrations and reliance on user-defined assumptions and models for behaviours of the aberrations to image quality (e.g. parabolic relationship). An image quality metric must also be defined that is dependent on the application and desired result.

The iterative approaches to SAO aberration correction considered above represent only a part of the field. Other approaches to SAO have been described in the Literature that do not require iterations. For example, phase stable Optical Coherence Tomography (OCT) images contain phase information that may be used to directly extract the wavefront aberrations using digital or computational AO methods [16,17] to be applied to a DM [18]. These approaches generally require a static sample or high-speed OCT acquisition in order to meet the phase stability requirement and may not be directly applicable to applications such as fluorescence confocal imaging where the phase information is lacking.

Deep Neural Networks (DNN) have also been applied to SAO approaches. In [19,20] a DNN was used to estimate the Zernike mode aberrations from a single measurement of the PSF in a non-imaging configuration. Deep Reinforcement Learning (DRL) has shown promise in real-world systems control and for complex optimization spaces, such as video games or chess [21–23]. Generalized deep learning control models have been shown to accurately correct aberrations in SAO [24]. Using DRL, a Deep Deterministic Policy Gradient (DDPG) was able to improve correction speeds by a factor of ~9 in comparison to a stochastic parallel-gradient descent algorithm by using a phase screen to generate aberrations and an image sharpness metric calculated on the PSF [25]. While using the PSF is compatible with confocal imaging applications in thin samples, it may not be directly compatible with OCT imaging in thick samples.

The iterative SAO methods that are described in this report use a sharpness-based merit function that is calculated on the image. This approach is readily used with either fluorescence images or OCT images and, when combined with an appropriate imaging system and sample, permits depth-resolved aberration correction based on the features of interest. Building on the work done for algorithmic and model-based approaches to SAO optimization, DRL can also be combined with SAO when framed as a control problem. In order to apply DRL to iterative models of SAO, the action space can be defined as the coefficients of the selected modes to be corrected, and the observations and rewards defined as the image quality metric.

A brief outline of DRL is provided in the description of the methods. The DRL was trained *in silico* and then transferred to a prototype confocal fluorescent microscope for proof-of-concept demonstration. The iterative image-based SAO aberration correction performance using DRL was compared against a coordinate-search approach using a static phantom sample.

Biomedical Optics EXPRESS

2. Methods

SAO methods are essentially an optimization of an image-based merit function. To overcome the limitations of algorithmic control approaches, a machine learning approach can be taken. Instead of using user-defined models as our optimization policy, Reinforcement Learning (RL) can be used where a computer automatically derives the best-fit model. For applications in real-world control, RL can be wrapped around a DNN to create DRL. This is especially useful in problems without labeled data where the observation, action, or optimization space is too unwieldy or ill-defined for a more typical DNN approach. DRL has the added benefit of environment exploration as it is not only fed with a state-reward pair but also decides on its actions. This essentially enables it to experiment and test hypotheses in its environment. Hence, DRL has significant potential for integration with SAO as a general control method. A brief introduction to RL and the concept of state-action pairs is included in Supplement 1 Section 1.

Where conventional DNNs create a complex lookup table and RL a learned policy, DRL benefits from a table of learned policies which are fine-tuned by gradient descent. This enables DRL to learn a policy-based method of optimization for complex tasks in hidden representations. Downfalls of DRL are its inherent short-sightedness that can be caused by sparse rewards and inability to resolve temporal data (for example, resolving the direction of motion of a moving object from a still picture). Temporal-resolution is improved by including a replay or experience buffer where every experience and sample batches are stored by the inclusion of preceding observations with the current one.

Policy gradients are a policy-based method that enable the action space to be continuous and compensate for the limitations in alternative value-based methods (such as Q-learning). DDPG [23] is a policy gradient method that is a form of Temporal Difference (TD) Learning that uses a separate network for the actor and critic to create a continuous action space. The DDPG actor decides the policy, which subsequently takes the action, and a critic evaluates the policy based on the TD error [26]. The schematic for the DDPG agent used in this report is shown in Fig. 1 [27].

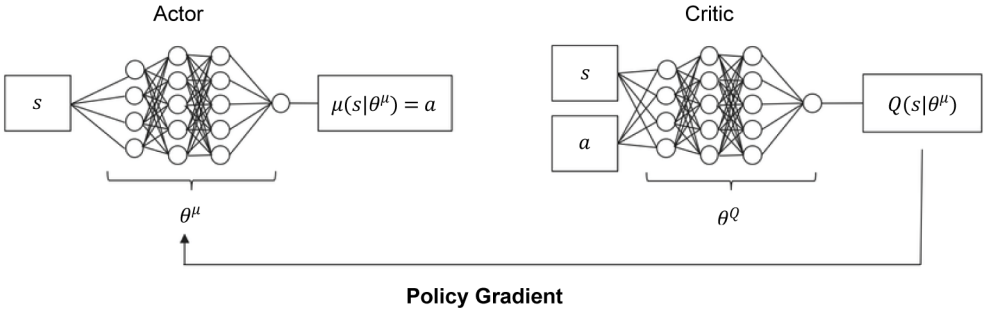

Fig. 1. DDPG schematic used in this report displaying actor-critic networks and their interaction via the policy gradient. A derivation of the policy gradient is provided in Supplement 1 Section 1.

To circumvent challenges with convergence in DRL, DDPG uses the same methods of experience replays and separate target networks as in Q-learning, but still falters with sparse rewards and local maxima problems. One method to correct for these challenges is to induce exploration in the agent by adding noise to the output action, essentially perturbing it and forcing it to take unchosen actions. An even more effective method is to induce noise in the parameter space, perturbing the decision-making process instead of the decision [28].

Python was used both *in silico* and *in situ* for the Machine Learning aspects of this project. The well-known library Keras with TensorFlow backend was used for the network architectures. Keras' DRL derivative, Keras-RL, was used for the DRL (DDPG) agent wrapper and logistics.

Biomedical Optics EXPRESS

The actor and critic networks were both simple fully dense nets with equivalent hidden layer structures as shown in Table 1 for 5 modes. Motivation for the simple architecture was its ease of use while prototyping solutions to a novel problem in order to eliminate simultaneous need of hyperparameter optimization and architecture exploration.

Table 1. Actor and Critic network architectures. *#modes* refers to the number of Zernike modes being corrected

	Actor Network		Critic Network	
Layer	Layer Dimension	Activation	Layer Dimension	Activation
Input	*Observation Space* (#modes*2 + 1)*(#modes+2)		*Observation Space + Action Space*	
Dense	500	ReLU	500	ReLU
Dense	1000	ReLU	1000	ReLU
Dense	1000	ReLU	1000	ReLU
Dense	1000	ReLU	1000	ReLU
Dense	1000	ReLU	1000	ReLU
Dense	500	ReLU	500	ReLU
Dense	300	ReLU	300	ReLU
Dense	200	ReLU	200	ReLU
Dense Output	*Action Space* (#modes)	tanh	1	Linear

Testing hyperparameters were set at learning rate of $1_E{}^{-5}$ with Adam optimizer and Mean Absolute Error loss for both networks. Rectified Linear Unit (ReLU) hidden layer activations with a tanh output activation were used for the actor network (to mimic the roughly zero-centered characteristics of Zernike mode coefficients) and the commonly used linear output activation was used for the critic network.

The agent's environment was self-programmed following OpenAI's Gym environment structure [29] where observations, actions, rewards, and reset conditions must be defined. The action space was defined as the coefficients of the selected Zernike modes while the observations and rewards were defined as the image quality metric. Initialization of the agent was coded into the environment as well as defining each step, which includes the agent taking an action, observing it and its associated reward, and updating its policy gradient. For consistency of terminology between DRL and the ZMHC methods, the observations are also referred to as steps.

Crafting a proper reward function is crucial for DRL convergence and behaviour, and thus also crucial for an SAO method to optimize the defined image quality metric. An example reward function is the image intensity squared and summed, assuming that intensity of the signal is maximum when the aberrations are corrected. Although normalizing the image quality metric may be desirable, zeroes can be introduced in denominators when dealing with zero-intensity signals, leading to artificially large merit function values that can confuse both SAO and DRL optimizations. For DRL, however, some normalization should be done to account for differing targets. Normalization using the maximum merit found within a set amount of training episodes was used instead of normalizing each image based solely on its own intensities.

2.1. Environment definition - observations

The observations were split into two categories – initial and dynamic. The initial observations included all modes' coefficients set to zero as well as the combination of each isolated mode's bounds – this contributes a space of $\#modes \cdot 2 + 1$ initial observations. These measurements can be considered to provide DeepMODAL with an initial low order fit to the observation space. The

Biomedical Optics EXPRESS

dynamic observations were continuously updated and served as "memory" of the most recent actions and rewards. These permitted DeepMODAL to refine the model of the optimization space and also provided updates for changes over time. Each observation included the combination of each mode's coefficient and the associated reward for that combination, leading to a total observation space of $(\#modes \cdot 2 + 1) \cdot (\#modes + 2)$. The quantity of dynamic observations was made a multiple of the number of modes but can be user-defined.

2.2. Environment definition - actions, rewards, and reset conditions

The action space included the entire continuous range of coefficients for the selected modes. Zernike mode coefficients were used due to simplicity and native support of the DMs for interpreting Zernike polynomials to actuator voltages.

The reward was based on the image quality metric defined by the image intensity squared and summed. As normalization is desirable to increase robustness to different imaging targets, signal powers, and imaging modalities, the rewards were normalized to a maximum image intensity metric in a defined number of steps. Due to noise influences and random fluctuations, non-linear normalizations were used to increase the difference between values around the maximum and other values. Normalization was done between [-1, 1].

Reset conditions are an important aspect of training where the conditions for starting a new episode are defined. Reset conditions should take into consideration that application using live targets will include motion, blinking, and artifacts. For training, the quicker of either 800 steps or reaching within 5% of the max value for more than 50 steps was taken as the reset condition. For implementation using live targets, constant recalibration via repeating initial observations would be detrimental to time and limit the effectiveness of the system – such applications should refrain from reset conditions unless the agent is unable to find a correction.

Several models for the DDPG network were rapidly prototyped and evaluated *in silico* as described in Section 2.3. Training of the final model was done using the physical system on a phantom for 200,000 episodes with the aberrating DM (as described in Section 3) generating the low order modes of defocus, horizontal coma, vertical trefoil, vertical astigmatism, and oblique astigmatism for a total of 5 aberrations in range of [-2.5, 2.5] μm *rms*.

2.3. In silico training environment

For prototyping, a simulation was used to model the system and aberration interactions as training *in situ* is time-limited by control of the DM actuators. During experimentation, training for 200,000 episodes took 4 hours *in silico* versus 20 hours *in situ*. To decrease overall training time, models can be first trained *in silico* then used in transfer learning or refined *in situ*. Our experiments showed that models trained *in silico* performed as well as those trained *in situ*.

The *in silico* model simulated realistic fluorescence imaging using the small animal cSLO imaging system used during validation as described in Section 3. A noise profile was defined as Poisson and additive half-normal noise due to photon and detector noises, respectively.

Wavefronts were calculated using the sum of Zernike modes given by their polynomial expansions [30,31]. The PSFs for the aberrating and correcting DMs were then superimposed and convolved with an image to derive the final aberrated image. The algorithmic equations can be found in Supplement 1 Section 2.

3. Imaging system

Training was done using a Dual-DM imaging system as schematically presented in Fig. 2. This system was a modification of a configuration used for imaging the mouse retina with SAO SLO [32] and OCT [33]. The SAO image-guided aberration corrections reported were performed on the fluorescence images only. Depth-resolved en face OCT images were subsequently acquired to demonstrate the multimodality capabilities of the system. The two DMs were optically conjugated

Biomedical Optics EXPRESS

to each other using relay lenses to match the diameter of each mirror. In these experiments, the sample was a phantom (lens cleaning tissue with fluorescent marker) and the DM-97 from AlpAO (DM_{AlpAO}) was used to create the aberrations. In the early proof of concept experiments, a simple VariOptics Variable Focus Lens was used in this place, generating one mode of aberration (defocus). The PTT-111 from IrisAO (DM_{IrisAO}) was controlled by DeepMODAL or the ZMHC method to correct the aberrations. A SH-WFS was placed behind a partially silvered mirror in a plane optically conjugated to the DMs. The two DMs played similar roles to acquire comparative results using the ZMHC approach. The DM_{IrisAO} served as the correcting element and the DM_{AlpAO} served the purpose of creating a range of aberrations, simulating the effects of differing aberration combinations. In a real application, this DM would be removed and the intrinsic aberrations arising from the sample would be corrected.

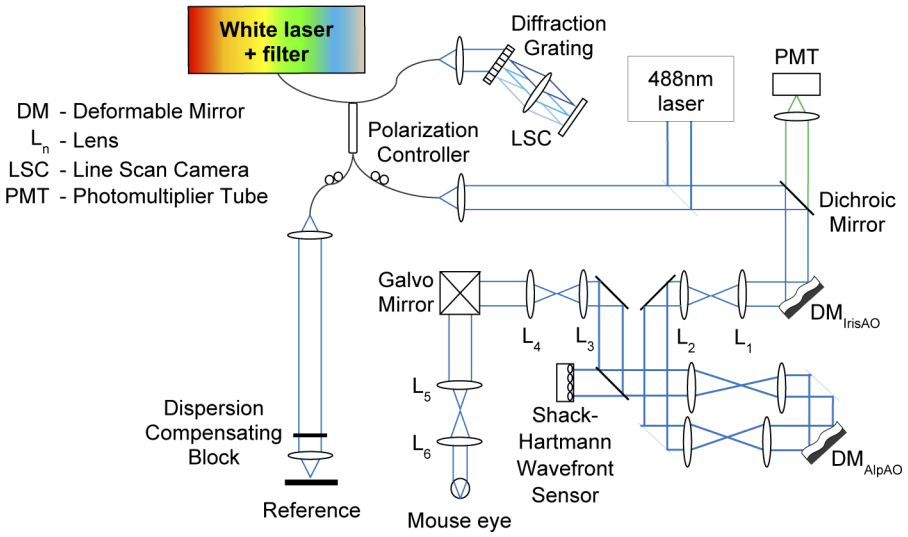

Fig. 2. Schematic of system.

4. Results

A DDPG framework, which was named Deep Multimodal Optimization by Direct Aberration Learning (DeepMODAL), was used as the DRL agent. The DDPG aberration-correction implementation was demonstrated for the five low-order Zernike modes of defocus, horizontal coma, vertical trefoil, vertical astigmatism, and oblique astigmatism with a continuous range of coefficients.

4.1. DeepMODAL - DDPG training

DDPG poses greater complexities and is prone to non-convergence and over-fitting due to sparse rewards on the continuous action space spectrum. The addition of parameter-space Gaussian white noise to perturb the decision-making process of the DDPG during training was used to 'force' exploration which was subsequently annealed, leading to convergence of an absolute maximal policy. Visualization of the training process and annealment of noise resulting in more consistent and accurate decision-making is displayed in Fig. 3.

Biomedical Optics EXPRESS

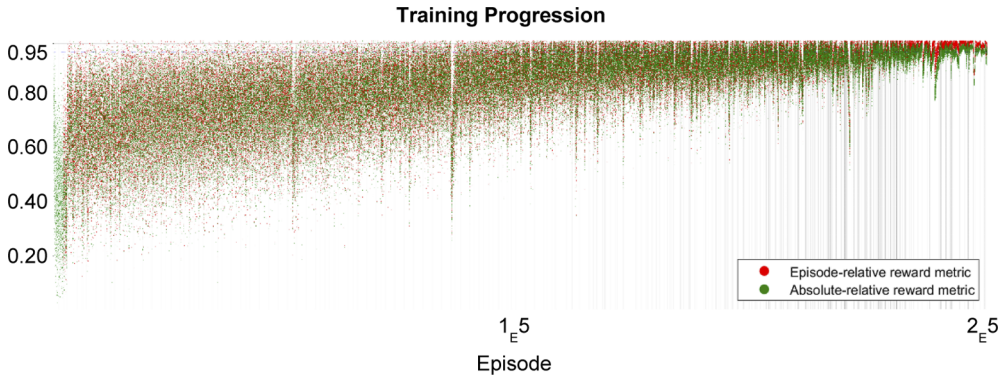

Fig. 3. Progression of agent training displaying high variation in the beginning due to noise-addition action-perturbation which is annealed as training progresses leading to high-accuracy convergence. Black vertical lines represent an aberration corrected, defined by an absolute-relative reward metric > .95, for at least 50 steps. Each dot represents a single time step.

4.2. *DeepMODAL - DDPG testing*

DeepMODAL was successfully able to correct for the 5 Zernike modes and noticeably improved image quality in less steps than a ZMHC approach. Figure 4 displays pre- and post-corrections using the DDPG agent. To check for robustness to artifacts, the sample was shifted vertically, longitudinally, and intermittent blocking of the optical path was introduced to mimic blinking. The agent was still able to correct given these realistic motions of a live imaging target due to movement and blinking, as visually and quantifiably shown.

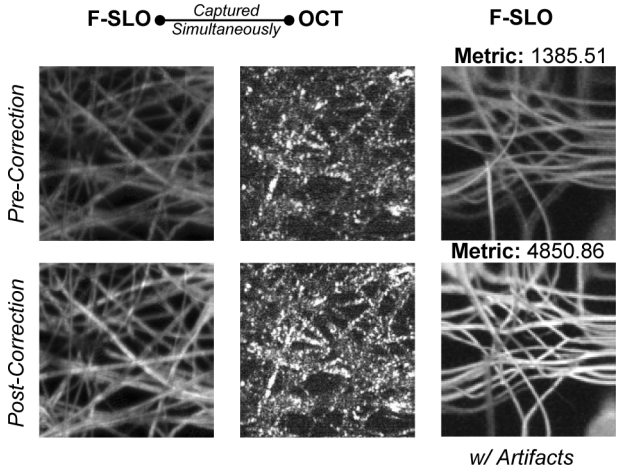

Fig. 4. Tissue phantom SLO aberrated by defocus, horizontal coma, vertical trefoil, vertical astigmatism, and oblique astigmatism (top) and corrected by DeepMODAL (bottom). The first two columns represent simultaneous images taken with a multimodal system whereas the third column shows another Fluorescence-SLO (F-SLO) acquisition of a sample that was physically shifted vertically, longitudinally, and blocked (mimicked blinking) during testing.

DeepMODAL optimizes in significantly less steps than ZMHC as shown in Fig. 5 and Fig. 6 (34 vs. 231 and 25 vs. 154 steps, respectively), and reaches a similar or better image quality metric even when optimizing for fewer modes. Figure 5 shows a representative result for the test

Biomedical Optics EXPRESS

case in which the aberrations generated by DM_{AlpAO} were within DeepMODAL's training bounds. After 34 initial + dynamic steps, or observations, DeepMODAL reached an image quality merit function value that was slightly higher than the ZMHC result after optimizing 7 modes over three iterations, totaling 231 steps (7 *modes* · 11 *steps* · 3 *iterations*).

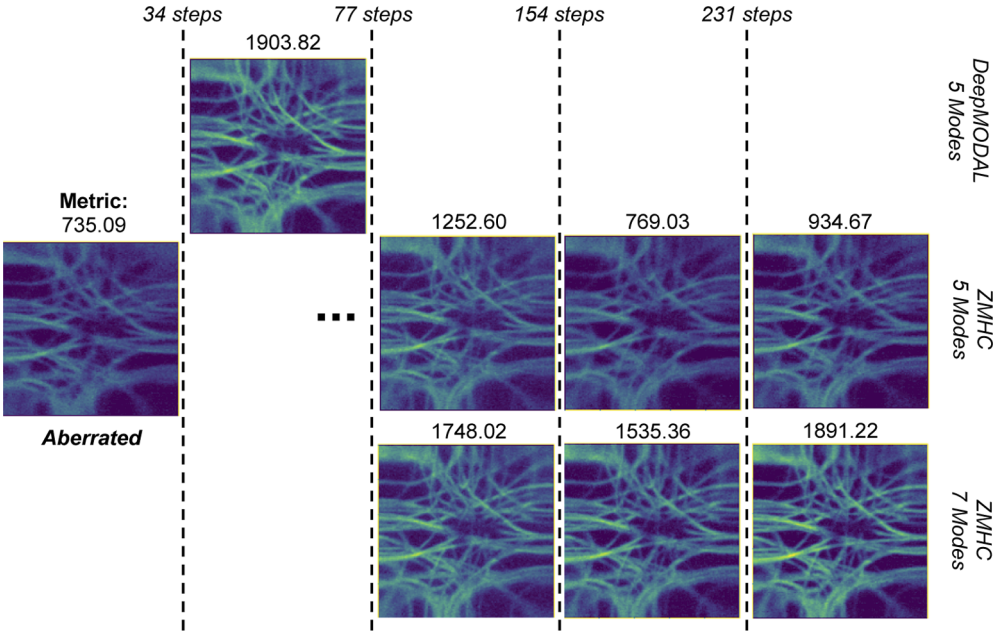

Fig. 5. Comparison of DeepMODAL (top) to ZMHC for 5 modes (middle) and 7 modes (bottom) with associated image quality metric. DeepMODAL's step count includes initial + dynamic steps.

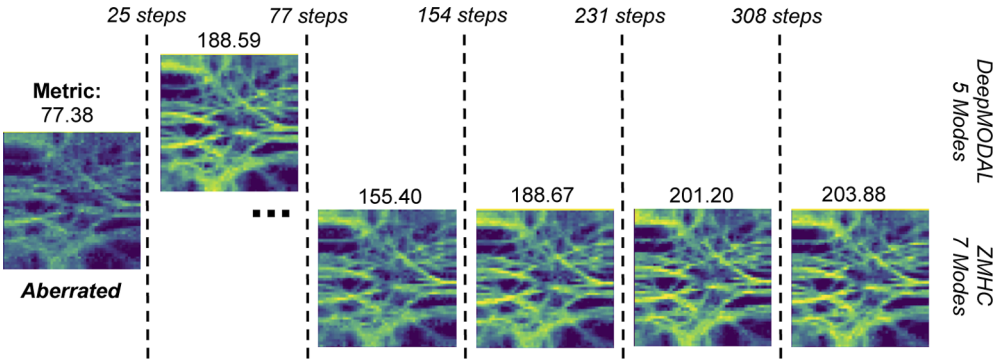

Fig. 6. Comparison for coefficient values outside of DeepMODAL's training bounds for DeepMODAL (top) to ZMHC for 7 modes (bottom) with associated image quality metric. DeepMODAL's step count includes initial + dynamic steps. Note: ZMHC eventually reaches a higher metric, but this is likely due to ZMHC correcting for more modes than DeepMODAL.

The middle row of Fig. 5 shows the images acquired with the results of the ZMHC method applied to the same 5 Zernike modes that DeepMODAL optimized by setting the coefficients for vertical coma and oblique trefoil to zero. After the second and third iterations of the ZMHC

Biomedical Optics EXPRESS

method, the value of the merit function calculated on the 5-mode images was significantly lower; this is due to the interaction of the modes during the ZMHC search [8,9] and demonstrates the ability for DeepMODAL to model the optimization function in a higher-dimensional space.

Figure 6 tests the ability of DeepMODAL to extrapolate by adding an amount of vertical astigmatism that was outside the training bounds. The ability to correct for extrapolated coefficients is significant as it suggests DeepMODAL creates an accurate inner representation of the aberration-reward relationship and optimizes it consistently, instead of merely creating a lookup table based on its training cases.

The ZMHC algorithm was run for 11 steps over 7 modes, leading to increments of 77 steps per iteration. Due to cross-talk of modes, ZMHC required multiple iterations over all modes [8,9]. It is important to note that in Fig. 6 the ZMHC was able to achieve an absolute higher image quality metric, though this may be due to its correction of higher-order Zernike modes to which DeepMODAL was not given access. Even with the smaller number of modes controlled, DeepMODAL can still perform similarly through optimization of the lower order modes. In either case, DeepMODAL outperforms ZMHC by a factor of 6-7 in terms of discrete steps with the added benefit of continuously-updating in the case of motion, whereas ZMHC would require new calibrations in the case of motion or artifacts.

4.3. *DeepMODAL's decision-making process visualized*

Meta-analyses of the testing progression reveal how DeepMODAL corrects for aberrations using SAO. Figure 7 shows a representative coefficient test progression during the dynamic steps phase. The first dynamic steps tend toward the ground truth and thereafter vary slightly around it. The amount of variance or exploration the agent does around the peak value can be tuned. Overly-aggressive variation of modal coefficients is undesirable when imaging live targets and attempting to image; however, slight exploration can improve the accuracy of finding a global maximum and makes the agent more robust to movement and artifacts. A realistic implementation would allow the agent to explore within reasonable bounds and "freeze" the

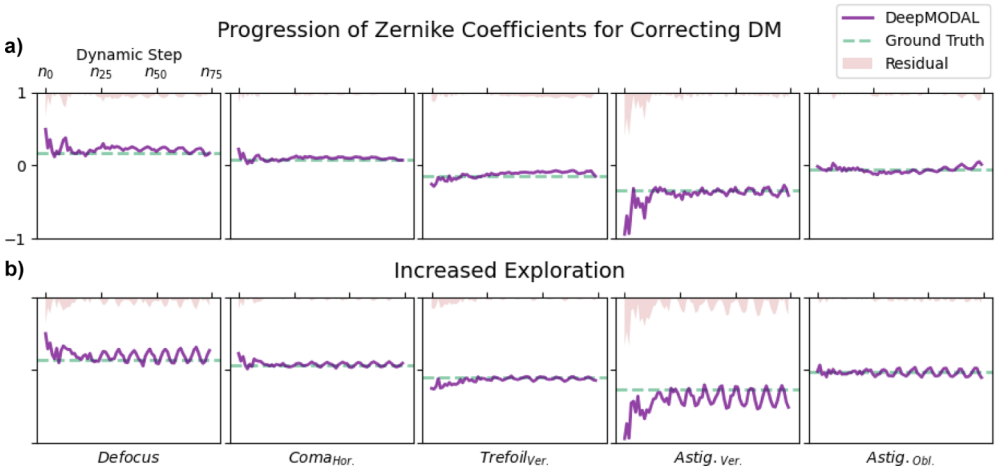

Fig. 7. Test case from Fig. 8 displaying progression of DeepMODAL's correction coefficients for 5 Zernike modes during the dynamic phase with associated ground truth and residual for standard implementation (a) and with increased exploration (b) as set by addition of parameter-space noise with amount being user-defined. The horizontal axes represent the dynamic step counts up to 75 steps. The vertical axes represent the Zernike mode coefficients in range [-1, 1] μm *rms*. The residual is represented in red in absolute terms from the top, i.e. $1 - |Residual|$.

Biomedical Optics EXPRESS

applied correcting aberrations based on the highest-found metric during image acquisition, where the agent's exploration could be temporarily paused and allowed to explore only while acquisition is not taking place. Another option is to cycle between exploration and acquisition based on time constraints and imaging protocols. Visualization of how increasing exploration effects the decision-making process of DeepMODAL is displayed in comparison of Fig. 7(a) (normal) versus Fig. 7(b) (increased exploration).

Figure 8 shows the images acquired during the representative test progression of DeepMODAL's initial and dynamic steps where the first image (I01) represents the pure intrinsically-aberrated image when the correcting DM (DM_{IrisAO}) is flat. The following initial steps represent the bounds of each mode individually isolated. The dynamic steps represent the exploration performed by the agent and the eventual convergence on a correcting solution. As can be seen, the agent will sometimes reach a high value and then a lower one. This is due to the agent probing the surrounding coefficient space, which is better visualized in Fig. 7.

Figure 9 displays the wavefront measured using a SH-WFS while Fig. 10 displays the associated Zernike mode decomposition during DeepMODAL's optimization as presented in Fig. 8 and Fig. 7(a). It can be seen that the wavefront flattens out, but slight deviations are expected due to SH-WFSs measuring non-common path aberrations due to their position laying outside the direct path of the camera and the target. The location of the SH-WFS is shown in the system schematic in Fig. 2.

Biomedical Optics EXPRESS

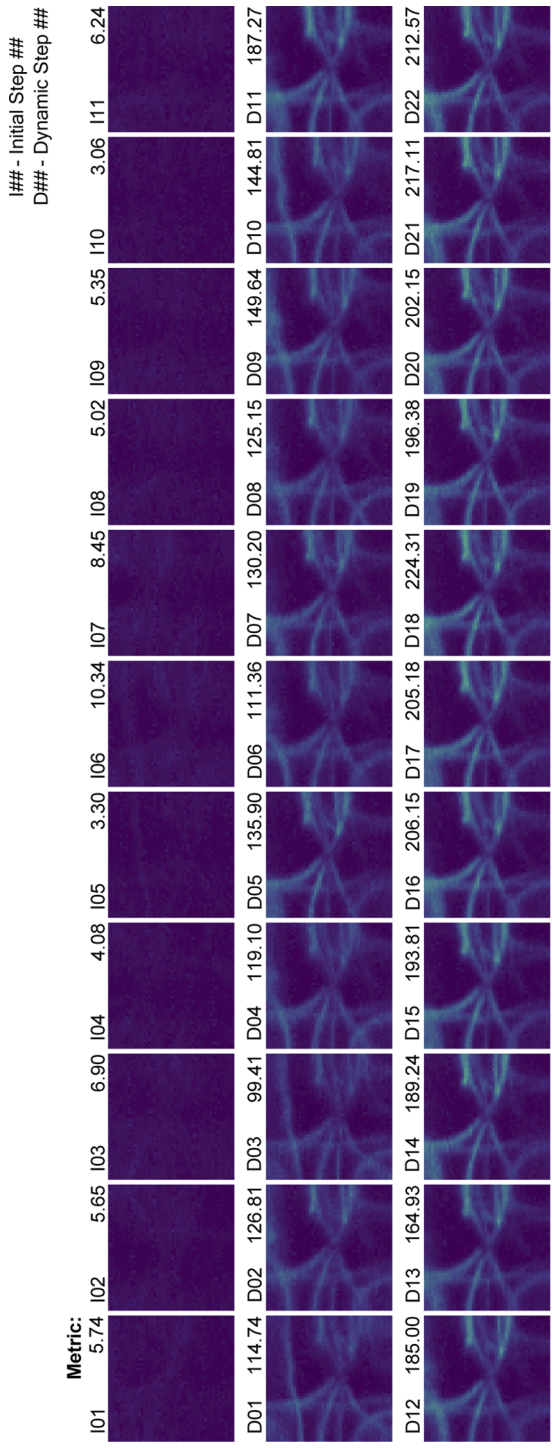

Fig. 8. Visualization of test case from Fig. 7(a) displaying progression of DeepMODAL's correction of a tissue phantom with associated image quality metric. I01 represents a flat correcting DM, i.e. the aberrated image, followed by the bounds of each of the 5 Zernike modes. D## represents the exploration and convergence of the agent.

Biomedical Optics EXPRESS

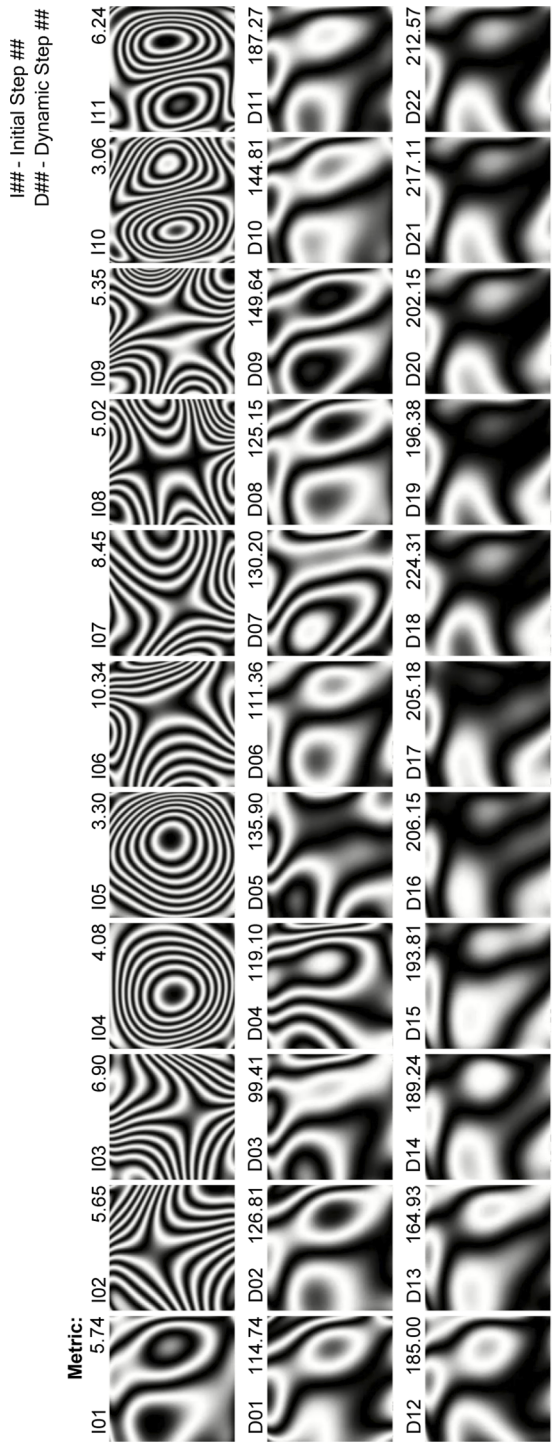

Fig. 9. Wavefront Sensor analysis of the test case from Fig. 8 and Fig. 7(a) displaying planar wavefront progression with associated image quality metric. I01 represents a flat correcting DM, i.e. the true aberration, followed by the addition of the bounds of each of the 5 Zernike modes. D## represents the exploration and convergence of the agent.

Biomedical Optics EXPRESS

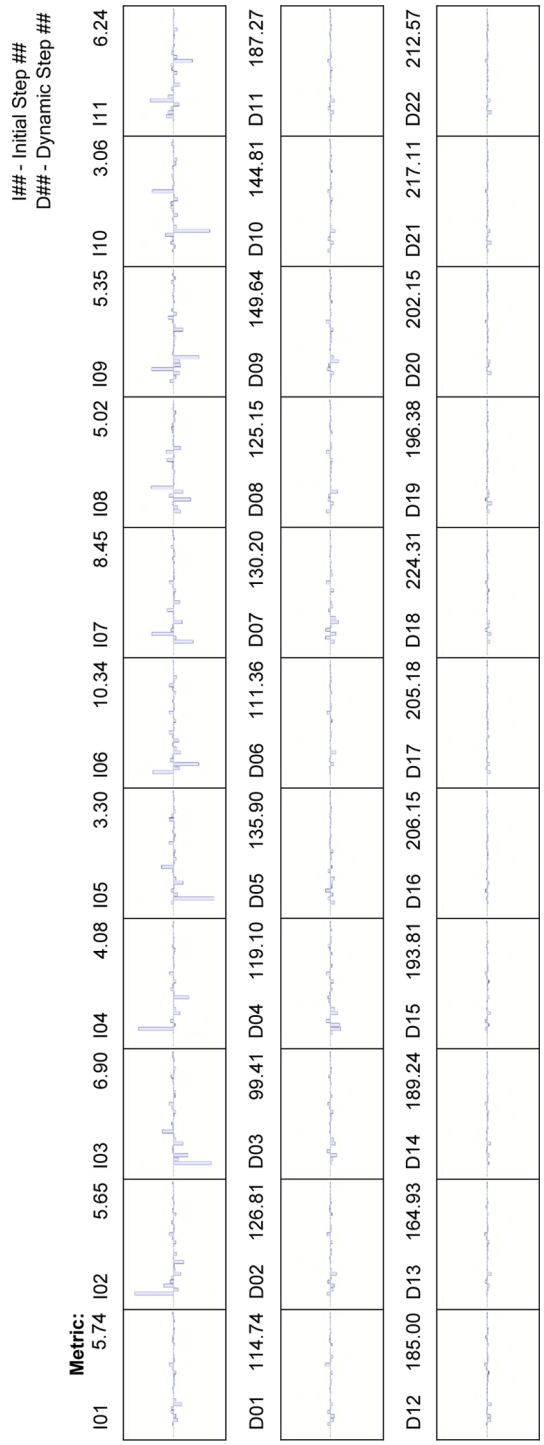

Fig. 10. Wavefront Sensor analysis of the test case from Fig. 8 and Fig. 7(a) displaying Zernike mode decomposition progression with associated image quality metric. The phase difference ranges from .3 to .13 λ. I01 represents a flat correcting DM, i.e. the true aberration, followed by the addition of the bounds of each of the 5 Zernike modes. D## represents the exploration and convergence of the agent.

Biomedical Optics EXPRESS

5. Discussion

By using Deep Reinforcement Learning to develop our SAO control algorithm, a computer-generated model was created to represent the relationship between observations and desired actions. In the case of SAO, our observations are a set of the most recent image quality metrics and our actions are the set of Zernike polynomial coefficients used to control the correcting DM. While SAO can be used to overcome non-common path aberration corrections and further optimize the image quality metric directly, its shortcomings lie in the time needed for correction as well as being limited by the accuracy of the correcting model. Using DRL seeks to overcome these problems by repeatedly refining a model until it accurately predicts the state-action relationship.

There are multiple policies that can be used in generating a model using DRL, one of the simplest being an extension of classical Q-learning to Deep Q Networks (DQN) [22]. DQNs, however, are limited by discrete action spaces which leads to actor-critic network policies such as DDPG as a possible solution to complex aberration-correction for large optimization spaces. DDPG features a continuous action space and thus wouldn't require high-dimensionality combinations to effectively cover the solution space. A DDPG was ultimately used in this project with fully-dense nets for both the actor and critic networks.

The agent can also be trained *in silico* to accelerate training and exploit greater processing power. The simulation proved to accurately model the problem; direct transfer of simulation-trained (*in silico*) models performed as well as real-world system trained models on real imaging targets without the need for further training. Simulation has a significant time overhead benefit which results in faster prototyping and training, as well as being able to exploit greater computational power, whereas a real-world system is generally limited by control and reaction times of the optical components (DMs, DAQ).

The success of transferring the results of the *in silico* training to the real-world case was partially due to the use of a sharpness based merit function to guide the DDPG. Our experiments showed that DeepMODAL was able to correct for different combinations of aberrations on different imaging targets. We speculate that DeepMODAL will be generalizable to other samples so long as appropriate imaging conditions are maintained. Artifacts in the images, such as spurious light, field of view larger than the isoplanatic patch, or low signal to noise ratio, would degrade DeepMODAL's performance.

Transfer-learning of the model is also possible to expedite training which may be necessary with increased complexity such as including a greater number of modes for correction. In future work, we anticipate increasing the number of modes corrected by DeepMODAL to 18 in order to match our previous work with the ZMHC approach. In such a case, more extensive training can be initially done *in silico* and refined on a real-world system. Such training approaches may also be necessary for complex architectures or agents which may require lengthy training. The transfer learning approaches may also be useful for future applications where DeepMODAL can learn to account for parameters such as DM hysteresis.

Architectures natively integrating time-series analysis may be necessary for an increased number of modes, as scaling of purely Dense networks may lead to an undesirable number of hidden connections which can confound and lengthen the learning process. Recurrent blocks such as Long Short-Term Memory (LSTM) [34] were tested and showed initial success on 5 and 7-mode correction using a single 128-dimension block, with an added benefit of training convergence in less steps. The large potential space to scale these Recurrent blocks and the severely reduced complexity further implies that Recurrent Neural Networks may be more applicable for scaling of DeepMODAL. Proper integration of memory within the network architecture instead of inputs can also forego or severely limit the dynamic step inputs, making the inputs a function of $\#modes$ instead of $\#modes^2$, further reducing complexity.

Biomedical Optics EXPRESS

6. Conclusion

This report demonstrated that the implementation of DeepMODAL using a DDPG can effectively correct for the 5 lower-order Zernike modes of defocus, horizontal coma, vertical trefoil, vertical astigmatism, and oblique astigmatism with 6-7 times less steps than a traditional hill-climbing coordinate search, ZMHC. Meta-analyses show how the agent probes around its initial guess and refines it to maximize a set image quality metric. Corrections are done by control of a Deformable Mirror employing Zernike mode coefficients, using only the image signal as input in a method known as Sensor-less Adaptive Optics. The agent is further able to explore the coefficient-metric space and continuously update, making it robust to real-world perturbations such as target movement, artifacts, and noise. The amount of exploration or variance from an optimal solution can also be user-set. Wavefront analysis also demonstrated that DeepMODAL effectively flattens the wavefront, but by direct control of every mode simultaneously as opposed to isolated modes as in ZMHC.

Funding. Natural Sciences and Engineering Research Council of Canada; Canadian Institutes of Health Research; Mitacs.

Acknowledgments. This research was supported in part by funding from NSERC, CIHR, and Mitacs.

Disclosures. MVS: Seymour Vision (I).

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplemental document. See Supplement 1 for supporting content.

References

1. S. Marcos, J. S. Werner, S. A. Burns, W. H. Merigan, P. Artal, D. A. Atchison, K. M. Hampson, R. Legras, L. Lundstrom, G. Yoon, J. Carroll, S. S. Choi, N. Doble, A. M. Dubis, A. Dubra, A. Elsner, R. Jonnal, D. T. Miller, M. Paques, H. E. Smithson, L. K. Young, Y. Zhang, M. Campbell, J. Hunter, A. Metha, G. Palczewska, J. Schallek, and L. C. Sincich, "Vision science and adaptive optics, the state of the field," Vision Res. **132**, 3–33 (2017).
2. B. C. Platt and R. Shack, "History and principles of Shack-Hartmann wavefront sensing," J. Refract. Surg. **17**(5), 379 (2001).
3. Y. Jian, R. J. Zawadzki, and M. V. Sarunic, "Adaptive optics optical coherence tomography for in vivo mouse retinal imaging," J. Biomed. Opt. **18**(5), 056007 (2013).
4. Y. Liu, J. Ma, B. Li, and J. Chu, "Hill-climbing algorithm based on zernike modes for wavefront sensorless adaptive optics," Opt. Eng. **52**(1), 016601 (2013).
5. A. Facomprez, E. Beaurepaire, and D. Débarre, "Accuracy of correction in modal sensorless adaptive optics," Opt. Express **20**(3), 2598–2612 (2012).
6. D. Débarre, M. J. Booth, and T. Wilson, "Image based adaptive optics through optimisation of low spatial frequencies," Opt. Express **15**(13), 8176–8190 (2007).
7. A. Camino, R. Ng, J. Huang, Y. Guo, S. Ni, Y. Jia, D. Huang, and Y. Jian, "Depth-resolved optimization of a real-time sensorless adaptive optics optical coherence tomography," Opt. Lett. **45**(9), 2612–2615 (2020).
8. R. R. Iyer, Y.-Z. Liu, and S. A. Boppart, "Automated sensorless single-shot closed-loop adaptive optics microscopy with feedback from computational adaptive optics," Opt. Express **27**(9), 12998–13014 (2019).
9. Z. Xu, P. Yang, K. Hu, B. Xu, and H. Li, "Deep learning control model for adaptive optics systems," Appl. Opt. **58**(8), 1998–2009 (2019).
10. K. Hu, B. Xu, Z. Xu, L. Wen, P. Yang, S. Wang, and L. Dong, "Self-learning control for wavefront sensorless adaptive optics system through deep reinforcement learning," Optik **178**, 785–793 (2019).
11. T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," arXiv:1509.02971 (2015).
12. D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedings of the 31st International Conference on Machine Learning*, Beijing, 2014.
13. R. Liessner, C. Schroer, A. Dietermann, and B. Bäker, "Deep reinforcement learning for advanced energy management of hybrid electric vehicles," in *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, Funchal, 2018.
14. M. Plappert, R. Houthooft, P. Dhariwal, S. Sidor, R. Y. Chen, X. Chen, T. Asfour, P. Abbeel, and M. Andrychowicz, "Parameter space noise for exploration," arXiv:1706.01905 (2017).
15. G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI Gym," arXiv:1606.01540, 2016.
16. D. J. Wahl, Y. Jian, S. Bonora, R. J. Zawadzki, and M. V. Sarunic, "Wavefront sensorless adaptive optics fluorescence biomicroscope for in vivo retinal imaging in mice," Biomed. Opt. Express **7**(1), 1–12 (2016).

Biomedical Optics EXPRESS

17. M. J. Ju, C. Huang, D. J. Jian, Y. Wahl, and M. V. Sarunic, "Visible light sensorless adaptive optics for retinal structure and fluorescence imaging," Opt. Lett. **43**(20), 5162–5165 (2018).
18. V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," arXiv:1312.5602, 2013.
19. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation **9**(8), 1735–1780 (1997).
20. J. Polans, D. Cunefare, E. Cole, B. Keller, P. S. Mettu, S. W. Cousins, M. J. Allingham, J. A. Izatt, and S. Farsiu, "Enhanced visualization of peripheral retinal vasculature with wavefront sensorless adaptive optics OCT angiography in diabetic patients," Opt. Lett. **42**(1), 17–20 (2017).
21. T. DuBose, D. Nankivil, F. LaRocca, G. Waterman, K. Hagan, J. Polans, B. Keller, D. Tran-Viet, L. Vajzovic, A. N. Kuo, C. A. Toth, J. A. Izatt, and S. Farsiu, "Handheld adaptive optics scanning laser ophthalmoscope," Optica **5**(9), 1027–1036 (2018).
22. H. Hofer, N. Sredar, H. Queener, C. Li, and J. Porter, "Wavefront sensorless adaptive optics ophthalmoscopy in the human eye," Opt. Express **19**(15), 14160–14171 (2011).
23. H. R. G. W. Wahls, S. Verstraete, J. Kalkman, and M. Verhaegen, "Model-based sensor-less wavefront aberration correction in optical coherence tomography," Opt. Lett. **40**(24), 5722–5725 (2015).
24. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," Nature **518**(7540), 529–533 (2015).
25. H. R. G. W. Verstraete, M. Heisler, M. J. Ju, D. Wahl, L. Bliek, J. Kalkman, S. Bonora, Y. Jian, M. Verhaegen, and M. V. Sarunic, "Wavefront sensorless adaptive optics OCT with the DONE algorithm for in vivo HUMAN retinal imaging [Invited]," Biomed. Opt. Express **8**(4), 2261–2275 (2017).
26. F. V. Zernike, "Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode," Physica **1**(7-12), 689–704 (1934).
27. R. J. Noll, "Zernike polynomials and atmospheric turbulence," J. Opt. Soc. Am. **66**(3), 207–211 (1976).
28. D. J. Wahl, P. Zhang, J. Mocci, M. Quintavalla, R. Muradore, Y. Jian, S. Bonora, M. V. Sarunic, and R. J. Zawadzki, "Adaptive optics in the mouse eye: wavefront sensing based vs. image-guided aberration correction," Biomed. Opt. Express **10**(9), 4757–4774 (2019).
29. V. Akondi and A. Dubra, "A two-layer Shack-Hartmann wavefront sensor model of the human and mouse retinas," in *SPIE BiOS*, Volume 11623, Ophthalmic Technologies XXXI (2021).
30. F. A. South, Y.-Z. Liu, A. J. Bower, Y. Xu, P. S. Carney, and S. A. Boppart, "Wavefront measurement using computational adaptive optics," J. Opt. Soc. Am. A **35**(3), 466–473 (2018).
31. B. Zhang, J. Zhu, K. Si, and W. Gong, "Deep learning assisted zonal adaptive aberration correction," Front. Phys. **8**, 634 (2021).
32. Y. Jin, Y. Zhang, L. Hu, H. Huang, Q. Xu, X. Zhu, L. Huang, Y. Zheng, H.-L. Shen, W. Gong, and K. Si, "Machine learning guided rapid focusing with sensor-less aberration corrections," Opt. Express **26**(23), 30162–30171 (2018).
33. A. Kumar, W. Drexler, and R. A. Leitgeb, "Subaperture correlation based digital adaptive optics for full field optical coherence tomography," Opt. Express **21**(9), 10850–10866 (2013).
34. Y. Geng, L. A. Schery, R. Sharma, A. Dubra, K. Ahmad, R. T. Libby, and D. R. Williams, "Optical properties of the mouse eye," Biomed. Opt. Express **2**(4), 717–738 (2011).